Econometría - ECN-323

Violación de Supuestos MRL: Heterocedásticidad

Francisco A. Ramírez
Economía - UASD

Abril, 2020

Violación de Supuestos del MRL

Violación de Supuestos del MRL: Introducción

- Caracterización empírica de la relación causal sugerida por la teoría económica es plasmada en el modelo de regresión lineal.
- Los supuestos son:

1. El modelo es lineal en los parámetros y el error entra de forma aditiva.

$$y_i = x_i'\beta + \epsilon_i$$

2. $E(\epsilon_i|x_i) = 0$
3. Homocedasticidad: $var(\epsilon_i|x_i) = E[\epsilon_i^2|x_i] = \sigma^2$

Violación de Supuestos del MRL: Introducción

4. No correlación serial (no autocorrelación): $E[\epsilon_i \epsilon_j | x_i] = 0$
5. No multicolinealidad: x_{ik} no son funciones exactas de otras variables
6. $\epsilon_i \sim N(0, \sigma^2)$

Violación de Supuestos del MRL: Introducción

- Estos supuestos permiten:

a) especificación de la función de regresión poblacional como una función lineal del conjunto de variables explicativas

b) obtención de un estimador de los parámetros desconocidos cuyas propiedades incluyen insesgamiento y varianza mínima (expresadas en el teorema Gauss-Markov), y

c) la posibilidad de hacer inferencia sobre los coeficientes estimados.

Violación de Supuestos del MRL: Introducción

- Ahora analizamos el impacto que tiene sobre la estimación e inferencia de los parámetros del MRL, la violación de los supuestos sobre los cuales se formuló el MRL.
- En particular, el análisis se realiza alrededor de tres preguntas:

1. ¿Cuáles son las consecuencias de la violación de un determinado supuesto sobre las propiedades del estimador de MCO?
2. ¿Cómo se puede detectar la inviabilidad o la violación de un supuesto?
3. ¿Cuáles alternativas existen para subsanar los efectos de la violación de uno de los supuestos del MRL?

Violación de Supuestos del MRL: Heterocedásticidad

Heterocedásticidad

- Hasta ahora hemos asumido que el valor promedio de la variable y es explicado por x a través de una función lineal

$$E(y_i|x_i) = x_i'\beta$$

- Para reconocer el impacto de otras variables no considerada, se incluye la variable ϵ_i para tomar en cuenta esos factores:

$$\epsilon_i = y_i - E(y_i|x_i) = y_i - x_i'\beta$$

- Por lo que el modelo para describir el comportamiento de y_i es:

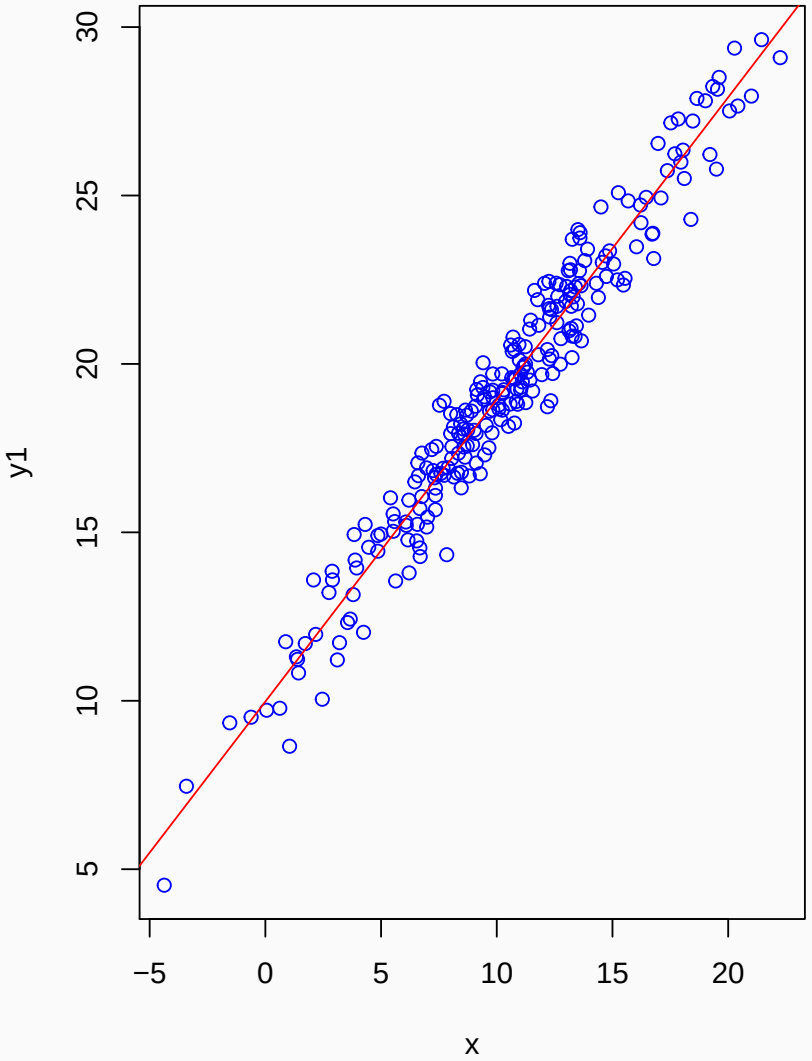
$$y_i = x_i'\beta + \epsilon_i$$

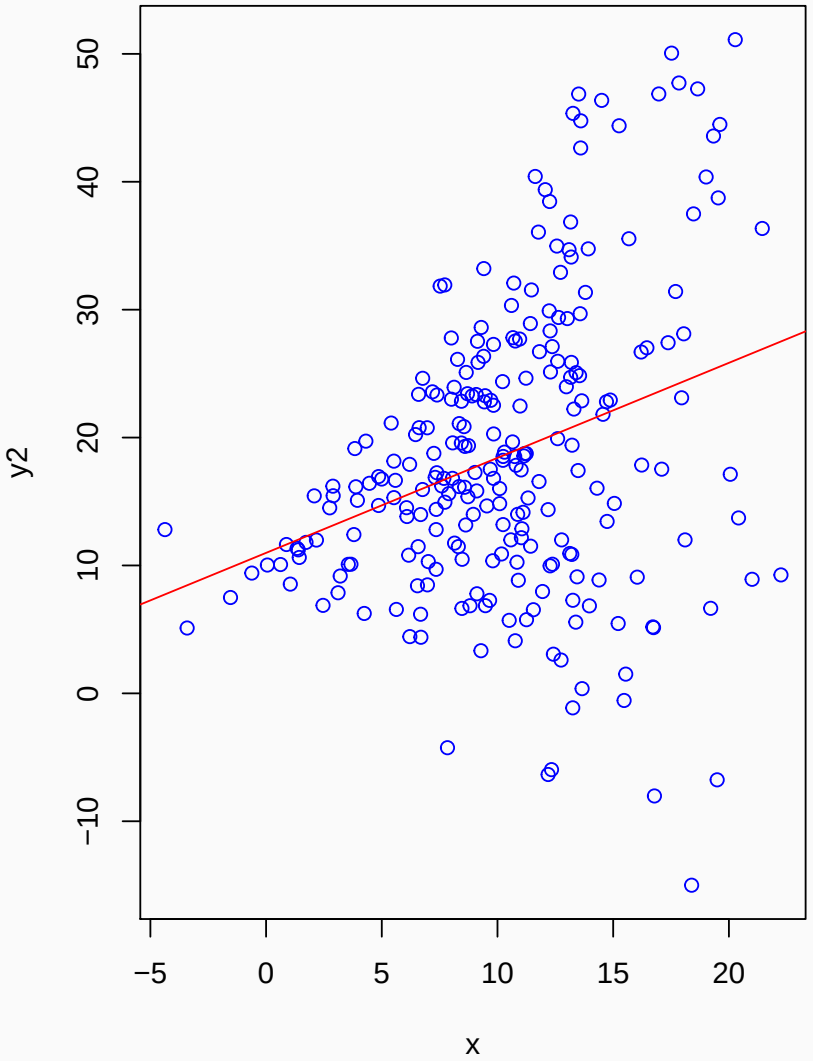
Heterocedásticidad

- Entre los supuestos en torno al modelo esta el de **homocedasticidad** o igual varianza para todos los ϵ_i.
- Hay situaciones donde este supuesto no es conveniente, pues puede llevar a conclusiones erradas de la relación entre las variables.

Heterocedásticidad

El gráfico siguiente muestra un ejemplo donde este supuesto es válido (izquierda) y uno donde no es válido (derecha).



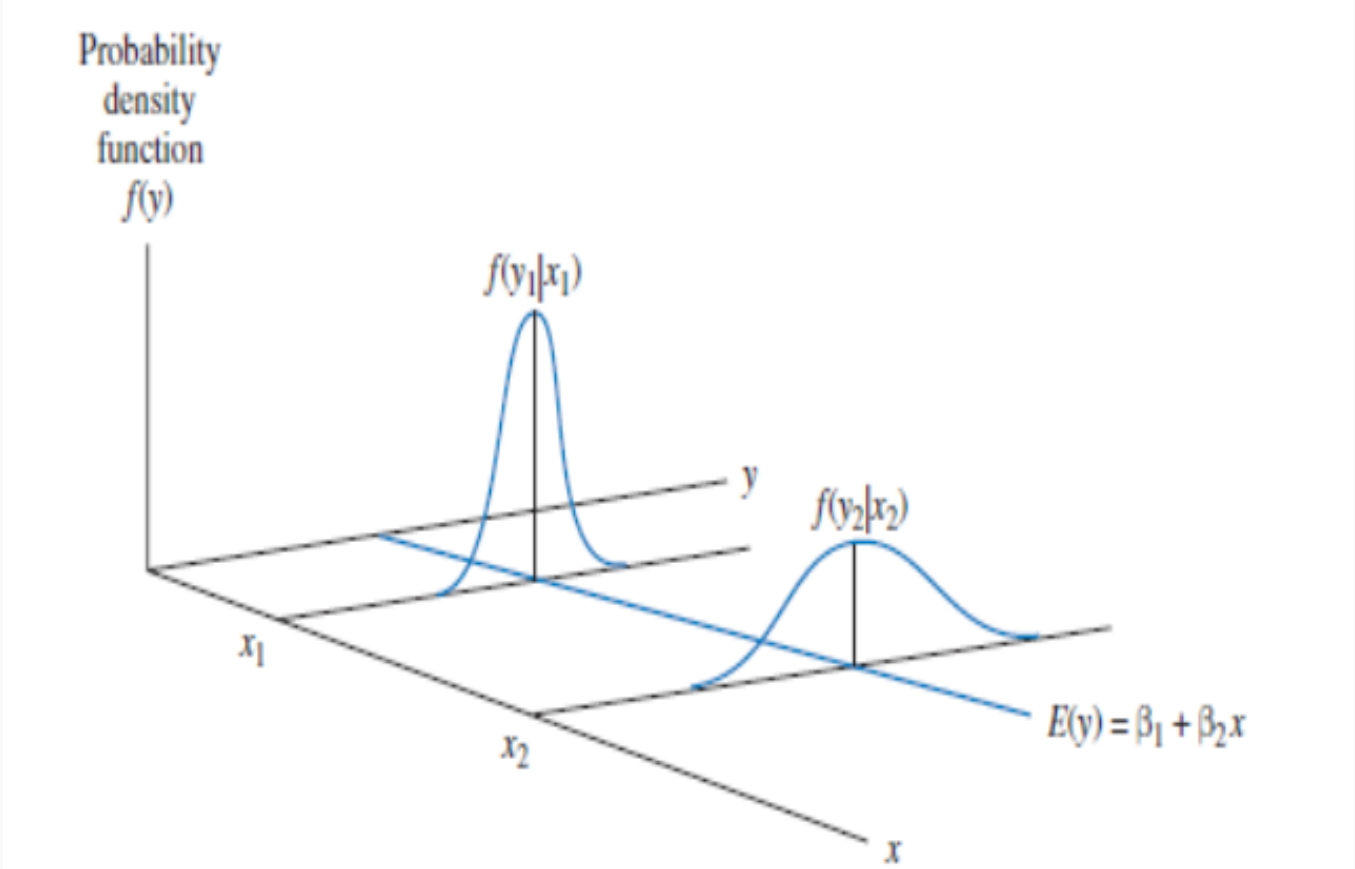


Heterocedásticidad

- Note que (en el gráfico de la derecha) se distinguen dos grupos de datos en esta muestra. El primer grupo tiene una dispersion menor que el segundo.
- En esta situación el supuesto de heterocedasticidad es inadecuado. Los errores del modelo se dice que en este caso son **heterocedásticos**.

Heterocedásticidad

- Gráficamente, lo que destaca es que la probabilidad de observar errores relativamente extremos es mayor entre distinto grupo de observaciones. En este gráfico ilustra el modelo de regresión lineal en presencia de heterocedasticidad.



Heterocedásticidad

- Se procede a reemplazar el supuesto:

$$var(\epsilon_i|x_i) = \sigma^2$$

por

$$var(\epsilon_i|x_i) = h(x_i)$$

- Donde $h(x_i)$ refleja que la varianza de ϵ_i es una función de x_i.
- Es importante destacar que la heterocedásticidad es un problema encontrado comunmente cuando se utilizan datos de corte transversal.

Consecuencias de la heterocedasticidad

- Los estimadores de MCO siguen siendo lineales e insesgados. Para ver esto note que:

$$b = (X'X)^{-1}Xy = \beta + (X'X)^{-1}X\epsilon$$

- Aplicando valor esperado:

$$E(b|X) = \beta + (X'X)^{-1}X'E(\epsilon|X)$$

- Dado el supuesto de exogeneidad, se tiene que:

$$E(b|X) = \beta$$

- Es decir, el insesgamiento no se ve afectado por la presencia de heterocedásticidad.

Consecuencias de la heterocedasticidad

- Sin embargo, la matriz de covarianzas (y por tanto los errores estándar), $var[b|X] = \sigma^2(X'X)^{-1}$ es la incorrecta. Para ver esto note que:

$$var[b|X] = E[(b-\beta)(b-\beta)'|X] = E\{[(X'X)^{-1}X'\epsilon][(X'X)^{-1}X'\epsilon]'|X)$$

$$= E\{[(X'X)^{-1}X'\epsilon][\epsilon'X(X'X)^{-1}]'|X)$$

$$= (X'X)^{-1}X'E(\epsilon\epsilon'|X)X(X'X)^{-1}$$

Consecuencias de la heterocedasticidad

- Definiendo como $E(\epsilon\epsilon'|X) = \Omega_\epsilon$ como la matriz de covarianzas, que tienen sobre la diagonal σ_i^2 (heterocedásticidad) para $i = 1, ..., N$ y ceros fuera de la diagonal (no correlación serial), se tiene que la matriz de varianzas de b es distinta al caso homocedástico.
- La implicancia es que los errores estándar usandos en la etapa de inferencia, es decir, en las pruebas de hipotesis e intervalos de confianza estan sesgados y toda la inferencia no es válida.

Consecuencias de la heterocedasticidad

- En el caso del modelo de regresión simple, puede apreciarse mejor el impacto de este problema:

$$b_2 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

$$var(b_2|x_i) = E[(b_2 - \beta_2)^2|x_i] = E\left[\left(\frac{\sum_i (x_i - \bar{x})\epsilon_i}{\sum_i (x_i - \bar{x})^2}\right)^2 \middle| x_i\right]$$

$$var(b_2|x_i) = E\left[\frac{\left(\sum_i (x_i - \bar{x})\epsilon_i\right)^2}{\left(\sum_i (x_i - \bar{x})^2\right)^2} \middle| x_i\right]$$

Consecuencias de la heterocedasticidad

- Analizando el numerador, tenemos:

$$E\left[\left(\sum_i (x_i - \bar{x})\epsilon_i\right)^2 \middle| x_i\right] = E\{[(x_1 - \bar{x})\epsilon_1 + ... + (x_N - \bar{x})\epsilon_N]^2 | x_i\}$$

$$= E\{[(x_1 - \bar{x})^2\epsilon_1^2 + ... + (x_N - \bar{x})^2\epsilon_N^2 + \textit{terminos cruzados}]|x_i\}$$

- Donde no se desarrollan los términos cruzados, que dado el supuesto de no correlación serial, son todos igual a cero cuando se aplica la esperanza matemática. Por tanto,

$$= (x_1 - \bar{x})^2 E(\epsilon_1^2|x_1) + ... + (x_N - \bar{x})^2 E(\epsilon_N^2|x_N) = \sum_i (x_i - \bar{x})^2 E(\epsilon_i^2|x_i)$$

Consecuencias de la heterocedasticidad

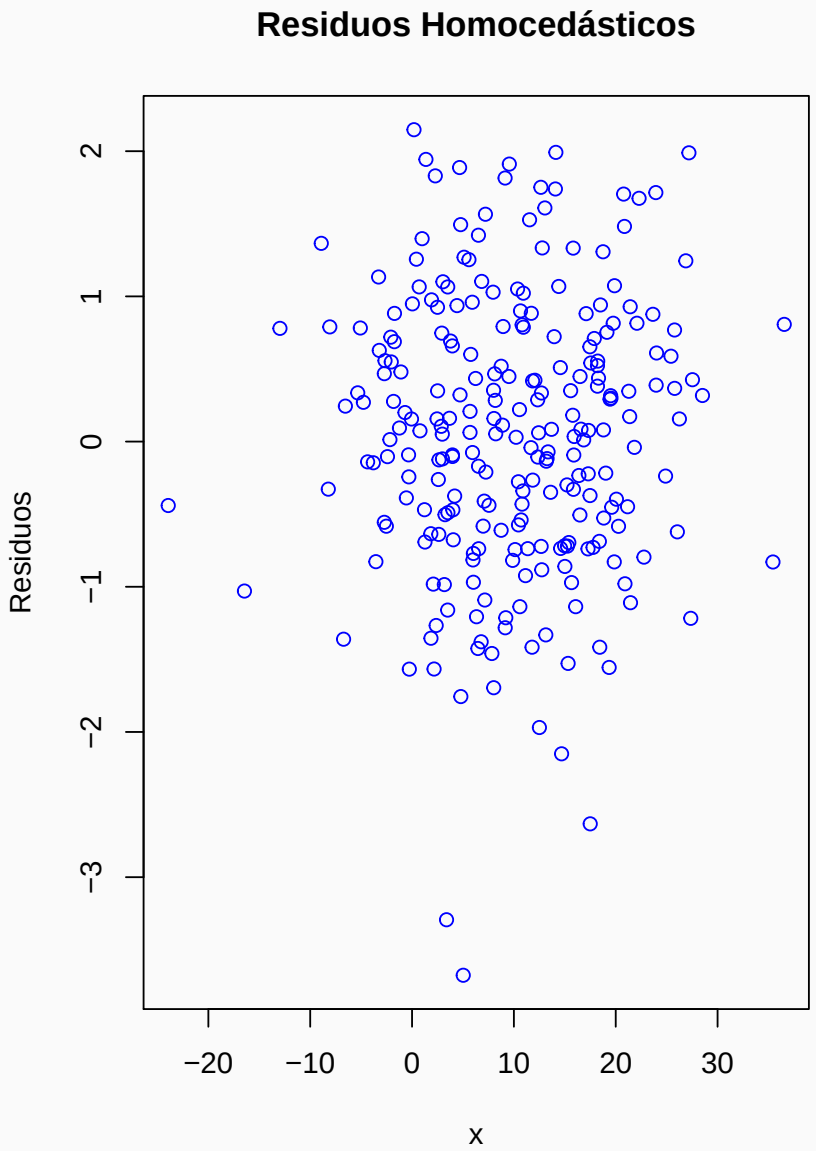
- Reemplazando se tiene que:

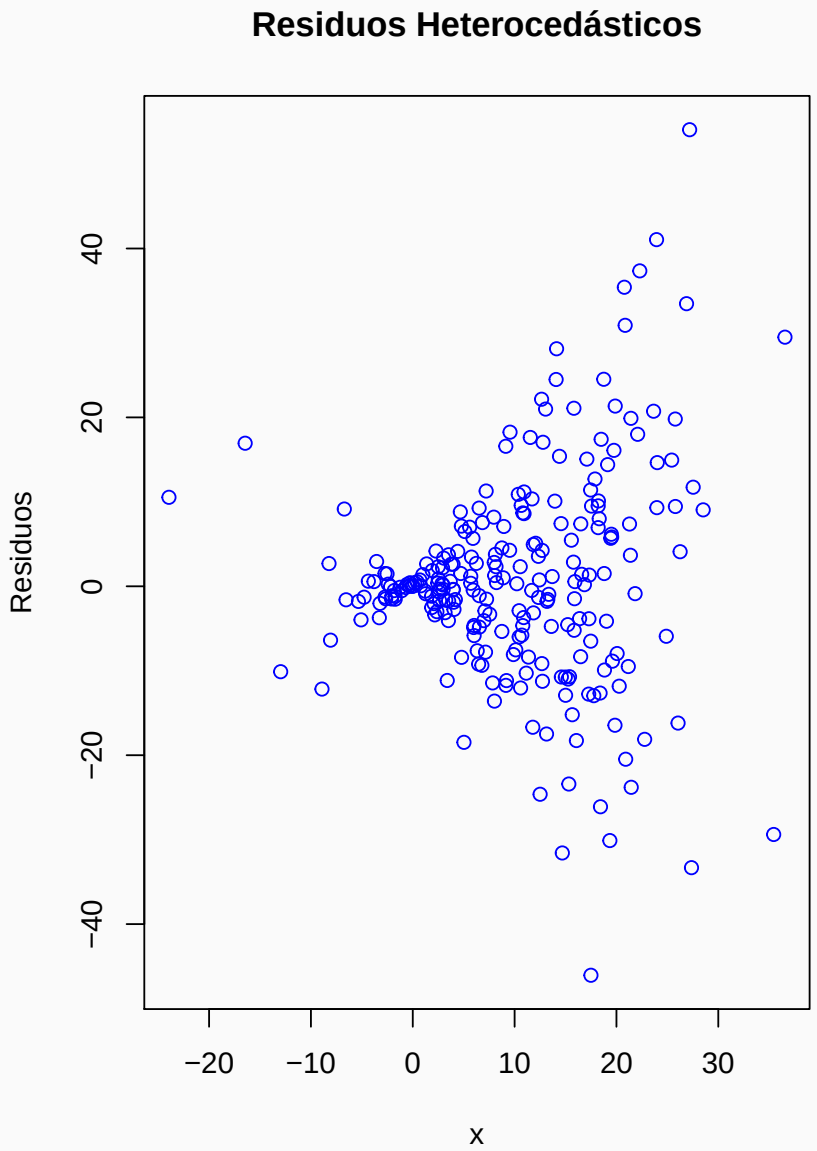
$$var(b_2|x_i) = \frac{\sum_i (x_i - \bar{x})^2 E(\epsilon_i^2|x_i)}{\left(\sum_i (x_i - \bar{x})^2\right)^2}$$

$$var(b_2|x_i) = \frac{\sum_i (x_i - \bar{x})^2 \sigma_i^2}{\left(\sum_i (x_i - \bar{x})^2\right)^2} \neq \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

Detectando la heterocedasticidad

Método 1: Análisis visual de residuos.





Detectando la heterocedasticidad

Método 2: Contrastes de Hipótesis

Contraste de Multiplicador de Lagrange o Contraste de Breusch - Pagan

- Defínase a la función de varianza de y_i como:

$$var(y_i|x_i) = \sigma_i^2 = E(\epsilon_i^2|x_i) = h(\alpha_0 + \alpha_1 z_{i1} + ... + +\alpha_s z_{is})$$

- Donde z_i puede ser igual o diferente que las x_i.

Detectando la heterocedasticidad

- Dos formas funcionales muy utilizadas para $h(.)$ son:

$$h(\alpha_0 + \alpha_1 z_{i1} + ... + + \alpha_s z_{is}) = exp(\alpha_0 + \alpha_1 z_{i1} + ... + + \alpha_s z_{is})$$

$$h(\alpha_0 + \alpha_1 z_{i1} + ... + + \alpha_s z_{is}) = \alpha_0 + \alpha_1 z_{i1} + ... + + \alpha_s z_{is}$$

- Cuando la varianza no depende de variable alguna, se obtiene el caso de varianza homocedástica:

$$h(\alpha_0 + \alpha_1 z_{i1} + ... + + \alpha_s z_{is}) = h(\alpha_0)$$

Detectando la heterocedasticidad

- En consecuencia, las hipótesis nula y alternativas para un contraste de heterocedásticidad basado en la función de varianza es:

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_s = 0$$

H_1 :No todos los α_j en H_0 son cero.

Detectando la heterocedasticidad

Para construir el estadístico de contraste, defina:

$$v_i = \epsilon_i^2 - E(\epsilon_i|x_i)$$

De tal manera que

$$\epsilon_i^2 = E(\epsilon_i|x_i) + v_i = \alpha_0 + \alpha_1 z_{1i} + ... + \alpha_s z_{si} + v_i$$

Como ϵ_i^2 es no observable, se sustituye por su estimado, el residuo:

$$e_i^2 = \alpha_0 + \alpha_1 z_{1i} + ... + \alpha_s z_{si} + v_i$$

Detectando la heterocedasticidad

- Estimando por MCO esta ecuación, se evalúa si las variables consideradas explican las variaciones en e_i^2. Utilizando el R^2 para medir la proporción de la varianza de ϵ_i explicada por las z_s, se tiene que si la H_0 es cierta el contraste se distribuye:

$$\chi^2 = N \times R^2 \sim \chi^2_{s-1}$$

Detectando la heterocedasticidad

Contraste de White

- Un problema con el contraste anterior es que presupone conocimiento de las variables en la función de varianza si la hipótesis alternativa es cierta.
- La propuesta de White es sustituir las z_s con las x_i, sus cuadrados y productos cruzados.
- Bajo las mismas hipótesis, el contraste de White es hecho bajo la prueba Chi cuadrada utilizando el contraste sobre el estadístico $\chi^2 = N \times R^2$.

Corrigiendo la heterocedásticidad.

- Como se observó anteriormente, la heterocedásticidad afecta la inferencia a través de su impacto sobre los errores estándar del los estimadores de MCO.
- La solución por lo tanto pasa por corregir dichos errores estándar para poder realizar la inferencia. Existen dos situaciones bajo las cuales dicha corrección puede ser implimentada:

a) la forma de la heterocedásticidad es conocida.

b) la forma de la heterocedásticidad es desconocida.

- En el primero de los casos el estimador resultante es conocido como Mínimos Cuadrados Generalizados o MCG, mientras que en el segundo como Mínimos Cuadrados Generalizados Factibles.

Corrigiendo la heterocedásticidad.

Estimador de MCG

- Cuando la forma de la heterocedásticidad es conocida la matriz de covarianzas de ϵ, $E(\epsilon\epsilon'|X) = \Omega_\epsilon$, es conocida.
- En ese caso, se transforma el modelo premultiplicando por $\Omega_\epsilon^{-1/2}$

$$\Omega_\epsilon^{-1/2} y = \Omega_\epsilon^{-1/2} X\beta + \Omega_\epsilon^{-1/2}\epsilon$$

- Definiendo: $y^* = \Omega_\epsilon^{-1/2} y$; $X^* = \Omega_\epsilon^{-1/2} X$ y $\epsilon^* = \Omega_\epsilon^{-1/2}\epsilon$, el MRL es:

$$y^* = X^*\beta + \epsilon^*$$

Corrigiendo la heterocedásticidad.

- El estimador de β es:

$$b_{MCG} = (X^{*\prime}X^{*})^{-1}X^{*\prime}y^{*} = (X'\Omega_{\epsilon}^{-1/2\prime}\Omega_{\epsilon}^{-1/2}X)^{-1}X'\Omega_{\epsilon}^{-1/2\prime}\Omega_{\epsilon}^{-1/2}y$$

- Dado que $\Omega_{\epsilon}^{1/2}$ es una matriz cuadrada y simétrica, se tiene que $\Omega_{\epsilon}^{-1/2}\Omega_{\epsilon}^{-1/2} = \Omega_{\epsilon}$ y que $\Omega_{\epsilon}^{-1/2\prime} = \Omega_{\epsilon}^{-1/2}$, se tiene que:

$$b_{MCG} = (X'\Omega_{\epsilon}^{-1}X)^{-1}X\Omega_{\epsilon}^{-1}y$$

Corrigiendo la heterocedásticidad.

Note que en este modelo transformado:

$$E(\epsilon^*\epsilon^{*\prime}|X) = E(\Omega_\epsilon^{-1/2}\epsilon\epsilon'\Omega_\epsilon^{-1/2\prime}|X) = \Omega_\epsilon^{-1/2}E(\epsilon\epsilon'|X)\Omega_\epsilon^{-1/2\prime}$$

$$= \Omega_\epsilon^{-1/2}\Omega_\epsilon\Omega_\epsilon^{-1/2\prime} = I$$

- Es homocedástica.

Corrigiendo la heterocedásticidad.

- Para fines ilustrativos, considere el siguiente modelo de regresión simple con heterocedásticidad:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

donde $E(\epsilon_i|x_i) = 0$, $var(\epsilon_i|x_i) = \sigma_i^2$, $E(\epsilon_i\epsilon_j|x_i) = 0$ para $(i \neq j)$.

- Donde se asume que:

$$var(\epsilon_i|x_i) = \sigma_i^2 = \sigma^2 x_i$$

Corrigiendo la heterocedásticidad.

- Como la varianza es conocida, se puede realizar una transformación del modelo:

$$\frac{y_i}{\sqrt{x_i}} = \beta_0 \frac{1}{\sqrt{x_i}} + \beta_1 \frac{x_i}{\sqrt{x_i}} + \frac{\epsilon_i}{\sqrt{x_i}}$$

$$y_i^* = \beta_0 x_{1i}^* + \beta_1 x_{i2}^* + \epsilon_i^*$$

- Para mostrar que $\frac{1}{\sqrt{x_i}}$ es la transformación adecuada, note que:

$$var(\epsilon_i^*|x_i) = E(\epsilon_i^{*2}|x_i) = E\left[\left(\frac{1}{\sqrt{x_i}}\epsilon_i\right)^2 \middle| x_i\right] = \frac{1}{x_i}E(\epsilon_i^2|x_i) = \frac{1}{x_i}\sigma^2 x_i = \sigma^2$$

Es decir, es homocedástica.

Corrigiendo la heterocedásticidad.

Estimador de MCGF

- En este caso se desconoce la forma exacta de la forma funcional de la varianza heterocedástica. Se parte asumiendo que la varianza depende de una o un conjunto de variables que explican el fenómeno.
- Se requiere una estimación de la matriz de covarianzas, $\hat{\Omega}_\epsilon$

Corrigiendo la heterocedásticidad.

- Para ilustrar este estimador, cosidere el caso del MRL simple. En este caso, asumiendo que la varianza depende de una o un conjunto de variables que explican el fenómeno, por ejemplo:

$$\sigma_i^2 = \exp(\alpha_1 + \alpha_2 z_i)$$

- Donde se utiliza la función exponencial para garantizar que la varianza es siempre positiva. Aplicando logatirmo natural:

$$\ln(\sigma_i^2) = \alpha_1 + \alpha_2 z_i$$

- Para la estimación de α_1 y α_2, se utiliza e_i^2 y se reescribe:

$$\ln(e_i^2) = \ln(\sigma_i^2) + v_i = \alpha_1 + \alpha_2 z_i + v_i$$

Corrigiendo la heterocedásticidad.

- La cual se estima por MCO, para obtener:

$$\hat{\sigma}_i^2 = exp(a_1 + a_2 z_i)$$

donde a_1 y a_2 son los valores estimados de α_1 y α_2.

Dividiendo el modelo por la varianza estimada:

$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \epsilon_i^*$$

donde $y_i^* = \frac{y_i}{\hat{\sigma}_i}$; x_{i1}^*; $x_i^* = \frac{x_i}{\hat{\sigma}_i}$.